CLASSIFIER-FREE GUIDANCE MAKES IMAGE CAPTIONING MODELS MORE DESCRIPTIVE

Anonymous authors
Paper under double-blind review

ABSTRACT

Image captioning is conventionally formulated as the task of generating captions that are similar to a set of human-generated reference captions, as measured using evaluation metrics such as CIDEr, ROUGE, and BLEU. Recent work has also explored reference-free captioning metrics based on the distance between generated captions and the corresponding images in the embedding space of a contrastively-trained image-text model such as CLIP. Here, we show that it is possible to trade off between reference-free and reference-based captioning metrics by decoding from a single autoregressive captioning model using classifier-free guidance (Ho & Salimans, 2021). Compared to standard greedy decoding, decoding from the same model with a guidance scale of 3 substantially improves caption→image retrieval performance when captions and images are embedded using CLIP (recall@1 49.4% vs. 26.5%) and CLIPScore (0.808 vs. 0.775), but greatly worsens standard reference-based captioning metrics (e.g., CIDEr 41.7 vs 126.1). Manual inspection reveals that higher guidance scales produce more descriptive but less grammatical captions.

1 INTRODUCTION

Image captioning is both a difficult task for computer vision systems to perform and a difficult task to evaluate. Although automated captioning metrics rank the best captioning systems higher than humans, human raters still show a strong preference for human-generated captions (Kasai et al., 2021b). Traditional maximum likelihood-based image captioning attempts to generate captions such that the $p(\text{caption}|\text{image})$ is high. However, captions from the ground truth distribution may describe only a subset of the salient aspects of an image, and are often non-specific, e.g., human annotators may describe a German Shepard only as a dog.

In this work, we explore classifier-free guidance (CFG) (Ho & Salimans, 2021) for image captioning on MS-COCO (Lin et al., 2014), finding that it yields an interesting trade-off in image captioning metrics. As we describe in greater detail below, CFG increases $p(\text{image}|\text{caption})$ at the expense of $p(\text{caption}|\text{image})$. Although CFG hurts "reference-based" image captioning metrics that measure the alignment between generated captions and human captions such as BLEU (Papineni et al., 2002), METEOR (Banerjee & Lavie, 2005), ROUGE (Lin, 2004), and CIDEr (Vedantam et al., 2015), it improves "reference-free" metrics that measure the similarity between the image and the generated caption in the embedding space of image-text models (Hessel et al., 2021). Qualitatively, we find that captions generated with CFG are more descriptive than both the ground truth captions and captions generated without CFG, but they are less grammatical, particularly at high CFG scales.

Related work. Early work using neural networks for image captioning found that models have a propensity to regurgitate captions from their training data, and as a result, the generated captions are not descriptive enough to uniquely identify images (Vinyals et al., 2015; Devlin et al., 2015). To address this shortcoming, Lindh et al. (2018) proposed to use an image retrieval model to examine whether images can be retrieved from generated captions, and to differentiate through this retrieval process to train a captioning model. Their approach marginally improves retrieval accuracy, but worsens reference-based captioning metrics. More recent work has adopted a similar approach to evaluate and improve the descriptiveness of captions based on the CLIP image-text model (Radford et al., 2021). Hessel et al. (2021) propose CLIPScore, an image captioning metric based on the cosine similarity between CLIP embeddings of the image and the generated caption. Kasai et al. (2021b) report that CLIPScore-based metrics align better with human judgments compared

to reference-based captioning metrics. Other work has jointly optimized CIDEr and CLIP-based losses using reinforcement learning, finding that the CLIP loss worsens standard reference-based captioning metrics but improves retrieval (Cho et al., 2022; Zhang et al., 2022).

Previous work has explored CFG for image captioning with diffusion models (Xu, 2022; Zhu et al., 2022) with little success. Xu (2022) report that it has little effect, whereas Zhu et al. (2022) observe an inconsistent benefit that varies with small changes in ratios of conditional to unconditional training examples. However, these studies investigate only reference-based captioning metrics.

2 METHODS

Let x be an image caption and y be the corresponding image. A standard captioning model aims to model the likelihood $p(\mathrm{x}|\mathrm{y})$, factorized autoregressively as $p(\mathrm{x}|\mathrm{y}) = p(\mathrm{x}_n|\mathrm{x}_{n-1}, \ldots \mathrm{x}_1, \mathrm{y}) \ldots p(\mathrm{x}_1|\mathrm{y})$ The network is trained so that $q_{\boldsymbol{\theta}}(\mathrm{x}_n|\mathrm{x}_{n-1}, \ldots \mathrm{x}_1, \mathrm{y}) \stackrel{\text{def}}{=} \mathrm{softmax}(f_{\boldsymbol{\theta}}(\mathrm{x}_{n-1}, \ldots \mathrm{x}_1, \mathrm{y}))$ approximates $p(\mathrm{x}_n|\mathrm{x}_{n-1}, \ldots \mathrm{x}_1, \mathrm{y})$, the probability of a token x_n given previous tokens $\mathrm{x}_{n-1}, \ldots, \mathrm{x}_1$. At inference time, it is common to use beam search or greedy decoding to produce a caption that has a particularly high probability.

Classifier-free guidance (CFG) (Ho & Salimans, 2021) aims to generate outputs that achieve a high value of $l_{\boldsymbol{\theta},\gamma}(\mathrm{x}, \mathrm{y}) \stackrel{\text{def}}{=} p(\mathrm{x})(p(\mathrm{x}|\mathrm{y})/p(\mathrm{x}))^{\gamma}$. Because $p(\mathrm{x}|\mathrm{y})/p(\mathrm{x}) = p(\mathrm{y}|\mathrm{x})/p(\mathrm{y})$ and $p(\mathrm{y})$ is fixed, $l_{\boldsymbol{\theta},\gamma}(\mathrm{x}, \mathrm{y}) \propto p(\mathrm{x})p(\mathrm{y}|\mathrm{x})^{\gamma}$, where γ is the classifier-free guidance scale. When $\gamma = 1$, the distribution is simply $l_{\boldsymbol{\theta},1}(\mathrm{x}, \mathrm{y}) = p(\mathrm{x}|\mathrm{y})$. Setting $\gamma > 1$ inflates the probability of the image given the caption $p(\mathrm{y}|\mathrm{x})$ relative to the unconditional probability of the caption $p(\mathrm{x})$.

Ho & Salimans (2021) originally proposed CFG in the context of diffusion models, which estimate the score functions $\nabla \log p(\mathrm{x}|\mathrm{y})$ and $\nabla \log p(\mathrm{x})$. Although $l_{\boldsymbol{\theta},\gamma}(\mathrm{x}, \mathrm{y})$ factorizes autoregressively, it is not a normalized probability distribution, so it is not entirely clear how one should sample tokens when performing autoregressive generation. Crowson (2022) suggested to sample from

$$\begin{aligned}\tilde{q}_{\boldsymbol{\theta},\gamma}(\mathrm{x}_n|\mathrm{x}_{n-1}, \ldots, \mathrm{x}_1, \mathrm{y}) \stackrel{\text{def}}{=} \mathrm{softmax}(&f_{\boldsymbol{\theta}}(\mathrm{x}_{n-1}, \ldots, \mathrm{x}_1, \mathbf{0}) \\ &+ \gamma(f_{\boldsymbol{\theta}}(\mathrm{x}_{n-1}, \ldots, \mathrm{x}_1, \mathrm{y}) - f_{\boldsymbol{\theta}}(\mathrm{x}_{n-1}, \ldots, \mathrm{x}_1, \mathbf{0})),\end{aligned} \tag{1}$$

where $f_{\boldsymbol{\theta}}(\mathrm{x}_{n-1}, \ldots, \mathrm{x}_1, \mathbf{0})$ are logits generated by the model without conditioning on the image. This formulation has been successfully applied in autoregressive image models (Yu et al., 2022b; Gafni et al., 2022). We adopt this formulation, but in our experiments, we decode greedily, i.e., at each step we take the token that maximizes $\tilde{q}_{\boldsymbol{\theta},\gamma}(\mathrm{x}_n|\mathrm{x}_{n-1}, \ldots, \mathrm{x}_1, \mathrm{y})$ and thus $l_{\boldsymbol{\theta},\gamma}(\mathrm{x}, \mathrm{y})$, so any choice of per-step normalization would yield the same captions.

Our model is a "bottleneck" variant of CoCa-Base (Yu et al., 2022a), which combines a contrastive loss with a captioning loss to simultaneously learn aligned image and text embeddings as well as a captioner. Whereas Yu et al. (2022a) condition the text decoder via cross-attention to pooled representations of the image encoder, this bottleneck variant uses only the contrastive image embedding. We adopt this bottleneck variant because of its simplicity and the conceptual appeal of producing a caption whose embedding lies close to the image embedding from the image embedding itself.

We pretrain our model on an image-text dataset comprising images from the JFT-3B dataset (Sun et al., 2017; Zhai et al., 2022) paired with their corresponding label names, web images paired with noisy alt text from the ALIGN dataset (Jia et al., 2021), and a small amount of data from other sources. We follow the same recipe as in Yu et al. (2022a), and do not mask conditioning information during pretraining. We then fine-tune on the MS-COCO train and Karpathy validation splits with a learning rate of 1×10^{-5} and vary guidance scale $\gamma \in \{1.0, 1.2, 1.5, 2.0, 3.0, 4.0\}$, conditioning masking proportion in $\{0.0, 0.25, 0.5, 0.75\}$, and numbers of steps in $\{5{,}000, 10{,}000, 20{,}000, 50{,}000\}$. Fine-tuning for $20{,}000$ steps with conditioning masking proportion 0.5 leads to near-optimal values of all metrics at all guidance scales; we report results for this model in tables.

We evaluate the standard reference-based captioning metrics CIDEr, METEOR, ROUGE, and BLEU-4, as well as two reference-free captioning metrics based on CLIP ViT-B/32 (Radford et al., 2021), on the MS-COCO Karpathy test split. The first reference-free metric is CLIPScore (Hessel et al., 2021), which is defined as $\texttt{CLIP-S}(\boldsymbol{c}, \boldsymbol{v}) = 2.5 \cdot \max(\cos(\boldsymbol{c}, \boldsymbol{v}), 0)$ where $\boldsymbol{c}$ and $\boldsymbol{v}$ are the CLIP embeddings of the caption and image respectively. The second reference-free metric is caption→image retrieval, measured as recall@k by taking the k images that are nearest to the generated caption in the CLIP embedding space and determining if the corresponding image is among

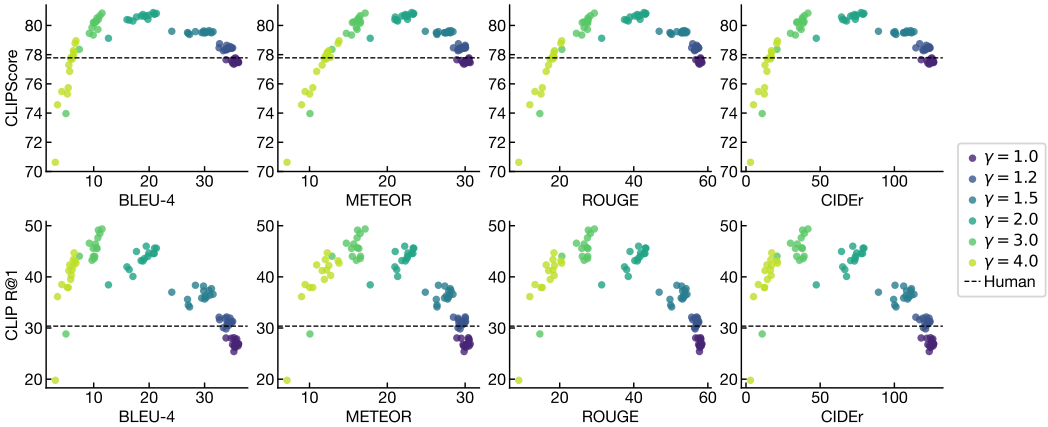

Figure 1: Performance of models trained with different hyperparameter combinations and evaluated with different guidance scales γ, as measured using reference-free captioning metrics based on CLIP ViT-B/32 (top: CLIPScore, bottom: recall@1) and reference-based caption metrics (BLEU-4, METEOR, ROUGE, and CIDEr). The dashed line reflects the value of the reference-free captioning metric for the ground-truth captions obtained from MS-COCO.

Table 1: Quantitative comparison of our approach with results from previous work that reports CLIP-based metrics, including UMT-BITG (Huang et al., 2021), VinVL-large (Zhang et al., 2021) (* indicates metrics from Kasai et al. (2021a)), CLIP-Captioner (Barraco et al., 2022), CLIP-S Reward (Cho et al., 2022), X-LAN+SCST+GEG (Zhang et al., 2022), and ZeroCap (Tewel et al., 2022).

	Reference-Based Metrics					Reference-Free Metrics				
Model	BLEU-4	METEOR	ROUGE	CIDEr	RefOnlyCLIP-S	CLIP-S	R@1	R@5	R@10	RefCLIP-S
Models trained with CLIP features or losses:										
CLIP-Captioner	38.7	29.3	58.6	126.0	0.811	0.754				0.814
UMT-BITG	37.3	28.2	57.9	122.6		0.772				
X-LAN+SCST+GEG	36.5	28.7	57.5	121.7			28.1	50.3	67.2	
CIDEr + CLIP-S Reward	37.7	28.8	58.3	124.6		0.772	24.4	50.2	63.1	
CLIP-S Reward	6.2	18.7	31.6	11.2		0.860	42.5	71.6	82.2	
ZeroCap	2.6	11.5		14.6		0.87				0.79
Models trained without access to CLIP:										
UMT-BITG w/o CLIP loss	37.6	28.3	58.1	122.5		0.725				
VinVL-large	**41.0**	**30.9**	**59.4***	**140.9**	**0.91***	0.78*				**0.84***
Ours ($\gamma = 1.0$)	36.1	30.5	58.2	126.1	0.900	0.775	26.5	51.9	64.1	0.830
Ours ($\gamma = 1.2$)	35.1	30.0	57.5	124.1	0.899	0.785	31.3	57.4	69.3	0.835
Ours ($\gamma = 1.5$)	31.5	28.4	54.4	113.2	0.891	0.796	36.6	64.0	75.0	**0.838**
Ours ($\gamma = 2.0$)	20.9	23.3	43.0	78.6	0.862	**0.808**	44.6	71.7	81.7	0.831
Ours ($\gamma = 3.0$)	11.5	17.1	29.4	41.7	0.820	**0.808**	**49.4**	**75.7**	**84.7**	0.811
Ours ($\gamma = 4.0$)	6.5	12.3	18.4	17.3	0.766	0.782	44.7	71.3	80.9	0.771

them. Because recall@k for $k > 1$ is highly correlated with recall@1 (R@5: $r = 0.99$, R@10: $r = 0.98$), we plot only recall@1. We additionally report RefOnlyCLIP-S, a reference-based metric based on the similarity of CLIP embeddings of the captions with embeddings of ground truth captions, and RefCLIP-S, which takes the average of the per-image harmonic means of CLIP-S and RefOnlyCLIP-S. Both RefOnlyCLIP-S and RefCLIP-S are discussed further in Hessel et al. (2021).

3 Results

Our main results concern the trade-off between reference-based and reference-free image captioning metrics as a function of guidance scale. Because different guidance scales and metrics could benefit from different fine-tuning hyperparameter combinations, we plot all results from our hyperparameter grid in Figure 1. Although standard greedy decoding ($\gamma = 1.0$) produces the highest CIDEr, ROUGE, BLEU-4 scores, higher guidance weights consistently yield higher values of reference-free captioning metrics. In particular, $\gamma = 3.0$ offers the highest caption→image recall and is tied with $\gamma = 2.0$ for the highest CLIPScore. Although we do not present them here, reference-free metrics calculated using CoCa image-text models exhibit similar patterns to CLIP ViT-B/32.

Table 1 compares our results, obtained from a single model evaluated at different guidance scales, with previous work that reports either CLIPScore or CLIP ViT-B/32 caption→image retrieval performance. Although our captioner is trained without access to CLIP and our pretraining dataset is distinct from CLIP's, sampling from our model with CFG yields higher CLIPScores than all other models trained without CLIP-based losses, and better CLIP caption→image retrieval even when compared with models that use CLIP-based losses.

γ=1.0: a car with a surfboard on top of it parked next to a car
γ=1.5: a vintage station wagon with a surfboard on top
γ=2.0: antique station wagons and a car buick stationwagon
γ=3.0: buick woody woody stationwagon and surf green station wagon parked in front of car show establishment
GT: An old green and brown car with chrome trim.

γ=1.0: a knife is sitting on a cutting board next to a sliced carrot
γ=1.5: a knife is sitting on a cutting board next to an orange
γ=2.0: knife sitting on cutting board next to whole one
γ=3.0: knife sitting on cutting board next to misshappi carrot on cutting board
GT: A knife sticking out of the side of a block of cheese.

γ=1.0: a herd of sheep grazing on a grass covered road
γ=1.5: sheep grazing on a highway with a truck on the road
γ=2.0: sheep graze on freeway medians where grass is grown
γ=3.0: grazing trucks blocking sheep on roadway grazing grass
GT: A large herd of sheep are grazing by the busy road.

γ=1.0: two birds standing in a cage with their heads in the air
γ=1.5: two birds standing inside of a cage in a zoo
γ=2.0: two crested cranes inside a wire cage
γ=3.0: crested tantalus cranes caged together in birdcage enclosure
GT: Two birds who are looking out of the cage they are in.

γ=1.0: a view of a city with a clock tower in the background
γ=1.5: a city with steeples and trees and buildings
γ=2.0: spires of churches line a city skyline
γ=3.0: spires steeples buildings trees church spires and trees
GT: A clock that is on the side of a tower.

γ=1.0: a kitchen with a microwave and a refrigerator
γ=1.5: a kitchen with a microwave and a refrigerator
γ=2.0: a kitchen with red appliances and white cupboards
γ=3.0: appliances sit in a small empty dingroomy red and white kitchen
GT: A kitchen that has a tile floor, a refrigerator, a microwave, and a toaster.

γ=1.0: a bathroom with a large mirror and a bathtub
γ=1.5: a bathroom with a large mirror and a bathtub
γ=2.0: a spacious bathroom with a large mirror and a large tub
γ=3.0: spacious bathroom with chandelier over tub mirrors and tv
GT: A bathroom with a tub, sinks, lights and a television.

γ=1.0: a vase filled with red and yellow flowers
γ=1.5: tulips in a clear vase on a table
γ=2.0: tulips in a clear glass vase on a tablecloth
γ=3.0: tulips in a clear punchov glass setting on doily
GT: Fresh red and yellow tulips in a vase.

Figure 2: Examples of captions generated with different classifier-free guidance scales, for randomly selected images from the test split.

To provide further intuition into the trade-off between reference-free and reference-based captioning metrics induced by CFG, we present examples of captions generated at different CFG scales in Figure 2. Higher CFG strengths lead to more descriptive captions. At $\gamma = 1.0$, the central object in the top left image is described as a "car" as in the ground truth caption, whereas at $\gamma > 1.0$ it is a "station wagon." Similarly, the animal in the top right image is described as a "bird" at $\gamma = 1.0$ and in the ground truth caption, but at $\gamma = 2.0$ becomes a "crested crane." However, higher CFG scales compromise grammaticality, leading to repeated words ("woody woody"), nonsense words ("misshappi", "dingroomy"), and lists of nouns arranged without verbs ("spires steeples buildings trees"). In several images, the captioning model describes what is present in the scene incorrectly, and CFG does not appear to substantially improve the output. The effect of CFG on reference-free metrics and descriptiveness cannot be explained solely by a change in caption length; as shown in Table 2, moderate CFG scales ($\gamma \leq 2$) have little impact on average numbers of words or characters per caption.

γ	Words	Characters
1.0	9.6 ± 1.4	44.2 ± 7.2
1.2	9.6 ± 1.4	44.7 ± 7.4
1.5	9.4 ± 1.4	45.7 ± 7.8
2.0	9.3 ± 2.4	50.3 ± 18.6
3.0	10.7 ± 7.6	69.0 ± 56.1
4.0	19.9 ± 16.9	161.2 ± 140.0

Table 2: Moderate CFG scales do not substantially change caption lengths, but higher scales result in longer captions. Numbers are mean $\pm$ standard deviation.

4 Conclusion

Our study demonstrates the utility of classifier-free guidance for generating descriptive image captions. At the same time, our results raise questions regarding the goal of image captioning and how it should be evaluated. As it is conventionally formulated, image captioning aims not to provide text that can substitute for an image, but to write the text that a human annotator would have written. This formulation penalizes captions that are more descriptive than ground truth, even when a human might prefer them. However, treating image captioning as a problem of generating a caption that lies close to the image in the embedding space of an image-text model is also inadequate, because captions that lie close to the image need not be grammatical and may contain gibberish. Yet our study suggests that modifying the decoding procedure can improve outputs (according to some criteria) even when the data are flawed. Although readily available large image-text datasets are noisy, there may nonetheless be ways to leverage these datasets to improve the trade-offs we describe.

REFERENCES

Satanjeev Banerjee and Alon Lavie. Meteor: An automatic metric for mt evaluation with improved correlation with human judgments. In *Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization*, pp. 65–72, 2005.

Manuele Barraco, Marcella Cornia, Silvia Cascianelli, Lorenzo Baraldi, and Rita Cucchiara. The unreasonable effectiveness of clip features for image captioning: An experimental analysis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4662–4670, 2022.

Jaemin Cho, Seunghyun Yoon, Ajinkya Kale, Franck Dernoncourt, Trung Bui, and Mohit Bansal. Fine-grained image captioning with CLIP reward. In *Findings of the Association for Computational Linguistics: NAACL 2022*, pp. 517–527, Seattle, United States, July 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.findings-naacl.39. URL https://aclanthology.org/2022.findings-naacl.39.

Katherine Crowson. You can apply a similar trick to classifier-free guidance to autoregressive transformers to sample from a synthetic "super-conditioned" distribution., 2022. URL https://twitter.com/RiversHaveWings/status/1478093658716966912.

Jacob Devlin, Hao Cheng, Hao Fang, Saurabh Gupta, Li Deng, Xiaodong He, Geoffrey Zweig, and Margaret Mitchell. Language models for image captioning: The quirks and what works. *arXiv preprint arXiv:1505.01809*, 2015.

Oran Gafni, Adam Polyak, Oron Ashual, Shelly Sheynin, Devi Parikh, and Yaniv Taigman. Make-a-scene: Scene-based text-to-image generation with human priors. *arXiv preprint arXiv:2203.13131*, 2022.

Jack Hessel, Ari Holtzman, Maxwell Forbes, Ronan Le Bras, and Yejin Choi. CLIPScore: A reference-free evaluation metric for image captioning. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pp. 7514–7528, Online and Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.emnlp-main.595. URL https://aclanthology.org/2021.emnlp-main.595.

Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. In *NeurIPS 2021 Workshop on Deep Generative Models and Downstream Applications*, 2021. URL https://openreview.net/forum?id=qw8AKxfYbI.

Yupan Huang, Hongwei Xue, Bei Liu, and Yutong Lu. Unifying multimodal transformer for bi-directional image and text generation. In *Proceedings of the 29th ACM International Conference on Multimedia*, pp. 1138–1147, 2021.

Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. Scaling up visual and vision-language representation learning with noisy text supervision. In *International Conference on Machine Learning*, pp. 4904–4916. PMLR, 2021.

Jungo Kasai, Keisuke Sakaguchi, Ronan Le Bras, Lavinia Dunagan, Jacob Morrison, Alexander R Fabbri, Yejin Choi, and Noah A Smith. Bidimensional leaderboards: Generate and evaluate language hand in hand. *arXiv preprint arXiv:2112.04139*, 2021a.

Jungo Kasai, Keisuke Sakaguchi, Lavinia Dunagan, Jacob Morrison, Ronan Le Bras, Yejin Choi, and Noah A Smith. Transparent human evaluation for image captioning. *arXiv preprint arXiv:2111.08940*, 2021b.

Chin-Yew Lin. ROUGE: A package for automatic evaluation of summaries. In *Text summarization branches out*, pp. 74–81, 2004.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*, pp. 740–755. Springer, 2014.

Annika Lindh, Robert J Ross, Abhijit Mahalunkar, Giancarlo Salton, and John D Kelleher. Generating diverse and meaningful captions. In *International Conference on Artificial Neural Networks*, pp. 176–187. Springer, 2018.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. BLEU: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pp. 311–318, 2002.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, pp. 8748–8763. PMLR, 2021.

Chen Sun, Abhinav Shrivastava, Saurabh Singh, and Abhinav Gupta. Revisiting unreasonable effectiveness of data in deep learning era. In *Proceedings of the IEEE international conference on computer vision*, pp. 843–852, 2017.

Yoad Tewel, Yoav Shalev, Idan Schwartz, and Lior Wolf. Zerocap: Zero-shot image-to-text generation for visual-semantic arithmetic. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 17918–17928, 2022.

Ramakrishna Vedantam, C Lawrence Zitnick, and Devi Parikh. CIDEr: Consensus-based image description evaluation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 4566–4575, 2015.

Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan. Show and tell: A neural image caption generator. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 3156–3164, 2015.

Shitong Xu. Clip-diffusion-lm: Apply diffusion model on image captioning. *arXiv preprint arXiv:2210.04559*, 2022.

Jiahui Yu, Zirui Wang, Vijay Vasudevan, Legg Yeung, Mojtaba Seyedhosseini, and Yonghui Wu. Coca: Contrastive captioners are image-text foundation models. *arXiv preprint arXiv:2205.01917*, 2022a.

Jiahui Yu, Yuanzhong Xu, Jing Yu Koh, Thang Luong, Gunjan Baid, Zirui Wang, Vijay Vasudevan, Alexander Ku, Yinfei Yang, Burcu Karagol Ayan, et al. Scaling autoregressive models for content-rich text-to-image generation. *arXiv preprint arXiv:2206.10789*, 2022b.

Xiaohua Zhai, Alexander Kolesnikov, Neil Houlsby, and Lucas Beyer. Scaling vision transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 12104–12113, 2022.

Pengchuan Zhang, Xiujun Li, Xiaowei Hu, Jianwei Yang, Lei Zhang, Lijuan Wang, Yejin Choi, and Jianfeng Gao. Vinvl: Revisiting visual representations in vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 5579–5588, 2021.

Youyuan Zhang, Jiuniu Wang, Hao Wu, and Wenjia Xu. Distincive image captioning via clip guided group optimization. *arXiv preprint arXiv:2208.04254*, 2022.

Zixin Zhu, Yixuan Wei, Jianfeng Wang, Zhe Gan, Zheng Zhang, Le Wang, Gang Hua, Lijuan Wang, Zicheng Liu, and Han Hu. Exploring discrete diffusion models for image captioning. *arXiv preprint arXiv:2211.11694*, 2022.